Associations and Study Design

Shannon Pileggi

STAT 217

The Data

From the CDC's 2013 Youth Risk Behavior Surveillance System

```
   gender height_m weight_kg  bmi carried_weapon bullied
1  female     1.73     84.37 28.2            yes      no
2  female     1.60     55.79 21.8             no     yes
3  female     1.50     46.72 20.8             no     yes
4  female     1.57     67.13 27.2             no     yes
5  female     1.68     69.85 24.7             no      no
6  female     1.65     66.68 24.5             no      no
7    male     1.85     74.39 21.7             no      no
8    male     1.78     70.31 22.2            yes      no
9    male     1.73     73.48 24.6             no     yes
10   male     1.83     67.59 20.2             no      no
...
8482 male     1.73     68.95  23             no      no
```


Response vs explanatory variable

In data analysis, we are generally interested in how the outcome or the **response** variable *depends on* or is *explained by* an **explanatory** variable.

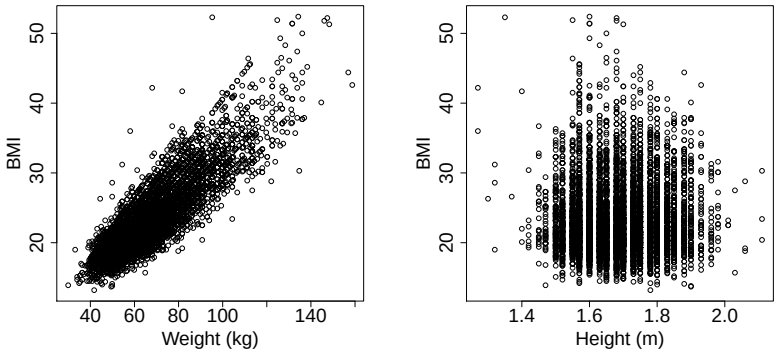
When there is a *relationship* between the two variables we say:

- there is an **association** between the response and an explanatory variable, or
- the response and an explanatory variable are **not independent**

When there is *no relationship* between the two variables we say:

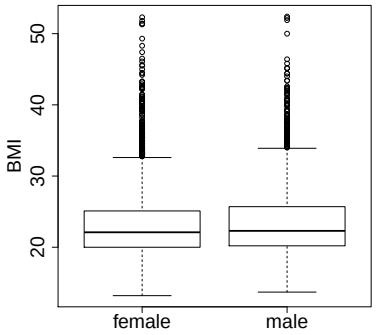
- there is **no association** between the response and an explanatory variable, or
- the response and an explanatory variable are **independent**

Two quantitative variables: Is there an association?



For descriptive statistics, we will use the correlation (to come later in the semester).

One quantitative and one categorical variable: Is there an association?

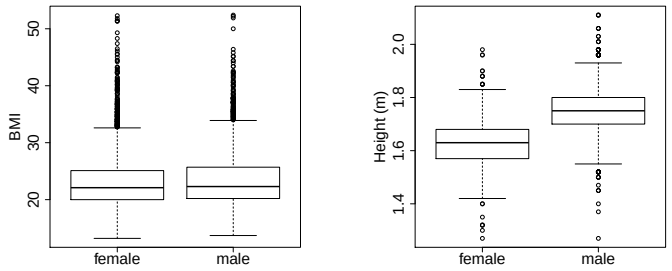


Group Exercise

Do you think there is an association between gender and bmi?

1. Yes, because the medians are different.
2. Yes, because the medians are the same.
3. No, because the medians are different.
4. No, because the medians are the same.

One quantitative and one categorical variable: Is there an association?



For descriptive statistics, we can report the mean and standard deviation in each group (or median and IQR). For example, the average height among males is 1.75 ± 0.08 m, and the average height among females is 1.62 ± 0.07 m.

Two categorical variables: Is there an association?

Carried Weapon	Males	Females	Total
Yes	1046	274	1320
No	3159	4003	7162
Total	4205	4277	8482

This is a 2x2 contingency table.

1. What percent of students carried a weapon to school?
2. What percent of students are male?
3. Among males, what percent carried a weapon to school?
4. Among females, what percent carried a weapon to school?
5. Among those who carried a weapon to school, what percent are male?
6. Which two percents should you compare if you want to know if gender can explain whether or not someone carries a weapon to school?

Interpreting a contingency table

Carried Weapon	Bullied: Yes	Bullied: Not	Total
Yes	312	1008	1320
No	1331	5831	7162
Total	1643	6839	8482

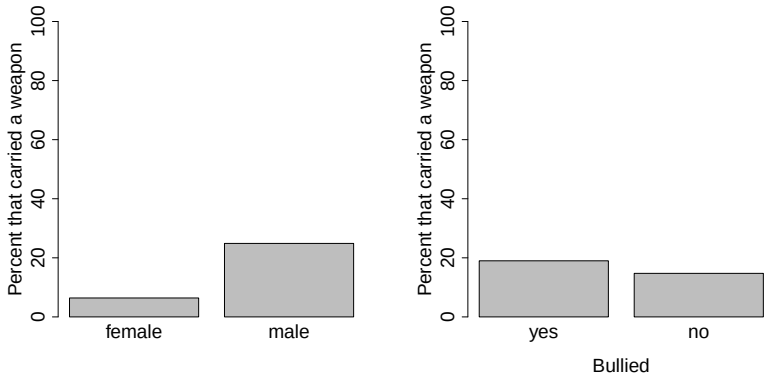
Which numbers should I compare in order to determine if there is an association between being bullied (explanatory variable) and whether or not a student carries a weapon to school (response variable)?

Group Exercise

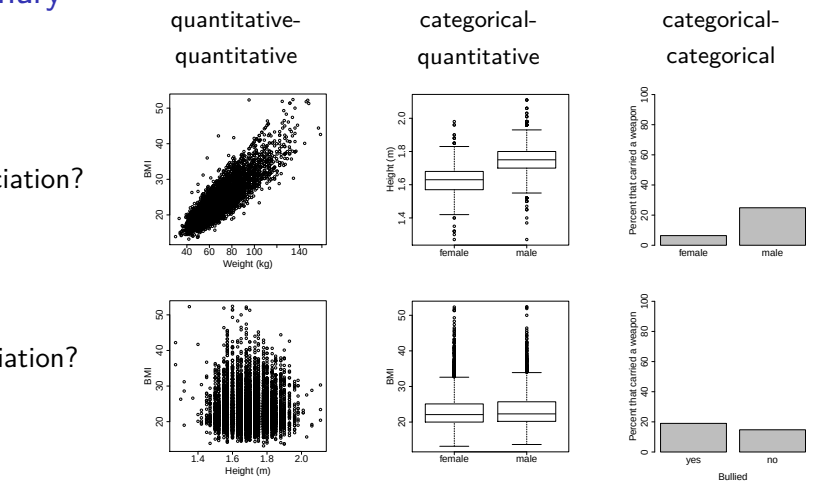
Which numbers?

1. 1643 vs 6839
2. 312 vs 1008
3. 312/1320 vs 1331/7162
4. 312/1643 vs 1008/6839
5. 1643/8482 vs 6839/8482
6. 1320/8482 vs 1643/8482

Two categorical variables: Is there an association?



Summary



We need formal statistical tests to determine the direction, magnitude, and significance of the association!

Associations between variables

Study Design

Types of studies

In an **experimental study** subjects are *assigned* to experimental conditions and then the response variable or outcome of interest is observed. The experimental conditions can be called **treatments**.

In an **observational study** researchers *observe* both the response and explanatory variable without assigning a 'treatment'. Observational studies are non-experimental.

We can study the effect of an explanatory variable on a response variable more accurately in an experimental study than an observational study.

Sampling discussion

Ideally, you want study participants to be a *representative* sample from your population so that your statistical inference can be *generalizable* to the population. Otherwise, your results may be *biased*.

Bias is present when the results of the sample are not representative of the population.

Potential sources of bias in observational studies

Sampling bias (or coverage bias) can result from the sampling method.

- Sample may not actually be random.
- The sample does not represent the entire population, resulting in **undercoverage** of certain groups in the population.

Nonresponse bias occurs when subjects refuse to participate.

- Participating subjects may have different characteristics than nonparticipating subjects.
- Participating subjects may choose not to response to some questions, generating **missing data**.

Response bias occurs when subjects give inaccurate answers.

- Subjects may lie.
- Question may be subjective or leading.

Group Exercise

Suppose I wanted to estimate the average GPA of all Cal Poly students. I use my STAT 217 class as a sample of all Cal Poly students.

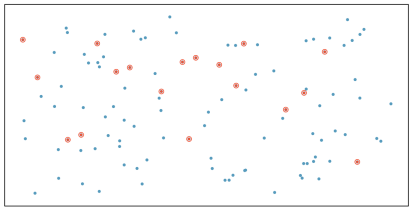
1. How could the following types of bias affect the study results?
   - sampling bias
   - nonresponse bias
   - response bias
2. Do you think the study results can be generalizable to all Cal Poly students?

Suppose we want to estimate the average age of college students, where college students are defined as individuals enrolled in higher education at community college (2 year institutions), traditional 4 year institutions, and online degree programs. We randomly select students from CalPoly and ask them their age.

Will the resulting average age of college students be biased? Will it overestimate or underestimate the average age of college students? Why?

1. unbiased because this would be a representative sample
2. biased due to response bias; average age of college students would be overestimated
3. biased due to sampling bias; average age of college students would be underestimated
4. biased due to non-response bias; average age of college students would be underestimated

Simple Random Sample



- Each individual equally likely to be samples
- Most likely to be *representative* of the population of interest (unbiased).

In order to conduct a simple random sample...

1. What do you need?
2. How do you do it?

Group Exercise

Investigators followed 806 kids age 2 to 4 and and 704 kids age 5 to 9 for four years. IQ was measured at the beginning of the study and again four years later. The researchers found that at at the end of the study the average IQ of kids who were not spanked was 5 points higher than spanked in the 2-4 group, and 2.8 points higher in the 5-8 group. The following newspaper headlines were observed:

- "Spanking lowers a child's IQ" (*Los Angeles Times*)
- "Do you spank? Studies indicate it could lower your kid's IQ" (*Houston Chronicle*)
- "Spanking can lower IQ" (NBC4i, Columbus, Ohio)
- "Smacking hits kids' IQ" (newscientists.com)

Based on the above information...

1. Is this an observational or experimental study?
2. Do you think these headlines accurately reflect the results of the study?

3 possible explanations

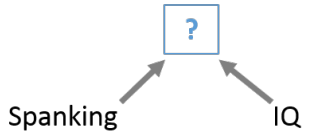
1. Spanking causes a decline in IQ

   Spanking (explanatory) → IQ (response)

2. Lower IQ causes kids to get spanked

   IQ (explanatory) → Spanking (response)

3. A *third* variable can explain both. A third variable that affects both the explanatory and the response variable and that makes it seem like there is a relationship between the two are called **confounding** variables.



In observational studies, association does not imply causation.

Headlines

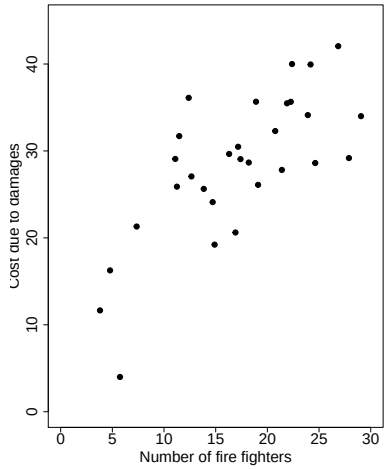
Incorrect interpretations:

- "Spanking lowers a child's IQ" (*Los Angeles Times*)
- "Do you spank? Studies indicate it could lower your kid's IQ" (*Houston Chronicle*)
- "Spanking can lower IQ" (NBC4i, Columbus, Ohio)
- "Smacking hits kids' IQ" (newscientists.com)

Correct interpretations:

- "Lower IQ's measured in spanked children" (world-science.net)
- "Children who get spanked have lower IQs" (livescience.com)
- "Research suggests an association between spanking and lower IQ in children" (CBSnews.com)

Group Exercise



Suppose we observe the relationship that more fire fighters are associated with more costs due to damages.

1. Does this figure mean that fire fighters *cause* damage?
2. What confounding variables could affect this relationship?

Key Concepts in Experimental Design

1. **Control** - compare treatment of interest to control group
2. **Randomize** - randomly assign subjects to treatment and control groups

Random assignment to treatment and control groups

Why randomly assign individuals to treatment and control groups?

1. comparing results between treatment and control groups actually allows us to determine if an intervention was effective
2. randomly assigning individuals to treatment and control groups allows us to make sure the groups are balanced with respect to other characteristics of the subjects
3. this allows us to attribute any observed differences as the result of the experimental assignment rather than confounding variables (can conclude a causal effect)

In a *well* designed experiment, results should not be affected confounding variables.

Types of studies

Experimental studies:

- reduces potential for confounding variables to affect results through random assignment
- may be able to conclude cause and effect
- may be unethical to assign 'treatment'
- typically has control and treatment group
- utilizes random assignment

Observational studies:

- confounding variables can affect the results
- cannot establish cause and effect
- may be easier to monitor a person's behavior
- typically has control and comparison group
- utilizes random sampling

Impact of study design on conclusions

	Random Assignment	Causation	Random Sampling	Generalizable
Ideal Experiment	✓	✓	✓	✓
Most Experiments	✓	✓	✗	✗
Most Observational	✗	✗	✓	✓
Weak Observational	✗	✗	✗	✗